Proceedings of the 10th International Aviation Psychology Symposium. Columbus, OH, 1999.

A MULTI-DIMENSIONAL EVALUATION METHODOLOGY FOR NEW COCKPIT SYSTEMS

Marcia Kuskin Shamo and Ravit Dror
Avionitek, Ltd
Haifa, Israel

Asaf Degani
Integral Human Systems
Palo Alto, California

ABSTRACT

It is essential that any system to be introduced to the cockpit for use by the flight crew be thoroughly evaluated. This evaluation must include a comprehensive range of human factors measures designed to provide a multi-dimensional assessment of the system in use. Additionally, the evaluation must be carried out within its unique operational environment. A rigorous assessment methodology is necessary both for the developers who design the systems, and for the regulatory agencies who must certify them as safe for airborne use. As an example of the multi-dimensional methodology, this paper describes the evaluation of a new cockpit system. We suggest that approach described here may be useful in defining a formal methodology for product development and certification.

INTRODUCTION

The need for a rigorous, comprehensive, and formal human factors evaluation methodology for advanced cockpit systems is well recognized (RTCA Task Force 4, 1999). A formal methodology is necessary for the aviation industry: designers must be able to assess not only the contribution of a new system to the performance of a specific task, but as well the interaction of the new system with other cockpit systems, and the impact of the new system on the general tasks of flight management. A methodology is also needed for the regulatory agencies who must develop viable requirements and standards for the effective human factors assessment and certification of these complex systems (RTCA Task Force 4, 1999; Abbott, Slotte, & Stimson, 1996). This methodology must provide a reliable means of evaluating the impact of the system on overall crew performance and on the safety of flight.

Although the human factors community has, over the years, developed a number of testing and evaluation concepts, methodologies, and tools for human factors assessment, their integration into a formal evaluation methodology has not occurred to any satisfactory degree (Stein & Wagner, 1994). Thus, there is no defined comprehensive

evaluation methodology which may be applied to the systems currently being developed for the cockpit.

No single evaluation methodology will, of course, be suitable for all new cockpit systems. Systems vary in the criticality of the functions they are intended to perform; criticality must have a major impact on the required comprehensiveness of the evaluation. Too, systems differ in the number and type of functions they execute, the amount and type of required human input, the level of integration with other systems, and so forth. Each of these system characteristics must be considered in the design of an evaluation methodology. While the exact mapping of evaluation methodologies to system characteristics is still largely undefined and far beyond the scope of this paper (see however Small & Rouse, 1994), we suggest that any system which plays a significant role in the cockpit must be submitted to a comprehensive, formal assessment. Further, this assessment will only be of value if it is carried out in the environment in which the system is intended to be used, and by the intended users.

This approach was utilized for the evaluation of a new computerized information system designed for the flight deck. The Integrated Crew Information System (ICIS) comprises individual units (one per crew member) that communicate via a local area network. The units contain a variety of procedural, operational, and informational applications that provide decision support for the crew (for example, normal and non-normal procedures, aircraft performance data and calculations, manuals, flight planning, moving maps, and charts). Task-oriented design enables immediate access to all information items relevant to the performance of each task. In addition, information retrieval is facilitated by multiple search mechanisms that allow the user to locate the desired item by name (index), physical location in the cockpit (a graphic virtual cockpit), or by using traditional tables of contents.

A formal and multi-dimensional evaluation of the ICIS was carried out at a major US airline. The goal of the Evaluation was to assess the ICIS on four dimensions considered relevant to the system and its intended role in the cockpit: safety, efficiency, usability, and acceptability. The Evaluation was conducted in two phases: Phase I was a full-mission evaluation which examined the use of the system within the operational environment, Phase II was a part-task evaluation in which the use of the system for specific tasks was examined in a controlled classroom setting. Phase II is reported in Shamo, Dror, & Degani, (1998); this paper describes Phase I of the ICIS Evaluation.

METHOD

Participants

Participants in the evaluation included line pilots and a number of support personnel:

Flight deck crew. Twenty-four Boeing 747-200 flight deck crews participated in the full-mission evaluation. Each crew included a captain (CA), a first officer (FO), and a second officer (SO), for a total of 72 participants.

Support personnel. Support personnel played an integral role in the ICIS Evaluation. The Ground Instructor (an instructor from the participating airline) recruited volunteers and

conducted the ICIS training. The airline's Simulator Facilitator was present in the simulator during the ICIS Evaluation session; he was responsible for operating the simulator and rating crew performance on standard airline rating forms. An Independent Expert Observer was present during the simulator sessions. The independent observer's task was to provide a non-biased expert rating of crew performance. In addition to rating the crew during the session and recording crew errors, the independent observer was responsible for providing the pre-session briefing, collecting data after the session, and crew error analysis.

Apparatus

ICIS units were installed in the cockpit for those crews who used the ICIS during the evaluation. All currently available paper manuals were present for those crews who did not use the ICIS during the evaluation.

Crews performed the scenario in a 747-200 full motion simulator. A single low light level (infrared) video camera was used to record crew actions and audio cockpit voice.

Tasks

All crews flew an identical mission scenario, which was created by several instructors of the participating airline. The scenario was designed to be representative of the normal and non-normal situations which might occur during a flight. The scenario, which started at the preflight checklist and ended with the aircraft back on the ground, required take-off speeds calculation and the calculation of abnormal landing speeds, normal and non-normal procedures execution (including multiple active non-normal checklists), fuel dumping, and an attempted engine restart. See Figure 1.

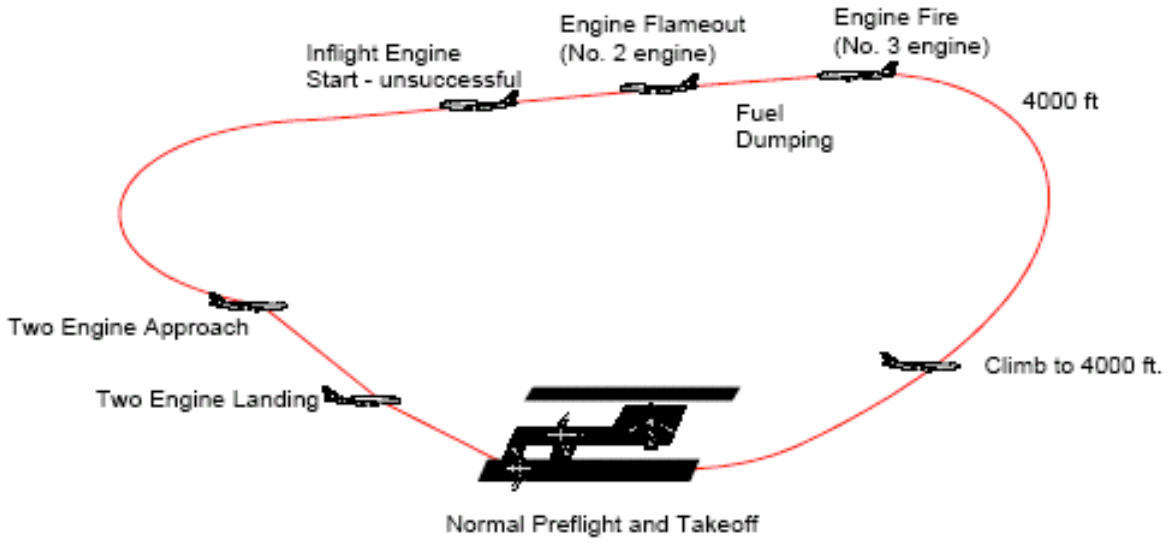

Figure 1. Mission Scenario

Measures and Tools

Measures and tools were defined for each of the four evaluation dimensions—safety, efficiency, usability, and acceptability:

- *Safety* was measured by crew error rate and severity (independent observer response, video analysis).
- Measures of the *efficiency* dimension included workload (NASA TLX), situational awareness (Situational Awareness Rating Technique – SART), crew resource management (simulator facilitator response to airline rating form, crew response), crew technical performance (simulator facilitator response to airline rating form, independent observer response, crew response), compliance with standard operational performance (independent observer response), and scenario execution times (video analysis).
- *Usability* and *acceptability* were assessed by subjective opinion questionnaires administered both after training and after the simulator session.

Experimental Design

The statistical design of the full-mission evaluation was a balanced between-subject design. Participants were randomly assigned to one of two groups:

- The control group performed a scenario in the simulator using paper procedures and manuals (CURRENT).
- The experimental group performed the same scenario using the ICIS only (ICIS).

Procedure

Training. All participants in the study, whether in the CURRENT or ICIS condition, received identical ICIS training, which lasted approximately 90 minutes.

Simulator Session. The simulator session took place two days after the training session. The independent observer briefed the crew and the simulator facilitator about the upcoming scenario, including the crew's assignment to ICIS or CURRENT condition. The simulator session then took place, with the independent observer present in the simulator. During the scenario the observer rated the crew and collected data on procedure execution. At the end of the session the crew completed the second set of questionnaires (TLX, SART, Post-Simulator opinion questionnaire); the independent observer then debriefed the crew. The simulator session (briefing, simulator, data collection, and debrief) lasted approximately 75 minutes.

RESULTS

The results are organized according to the four dimensions safety, efficiency, usability, and acceptability. For brevity only the major findings of the study are presented in detail below, others are summarized. Complete results are discussed elsewhere (Avionitek, 1998).

Safety

The independent observer recorded crew error rate during the simulator session. Subsequent to the session, video analysis of all simulator sessions allowed the observer to validate the recorded crew errors, list additional crew errors missed during the session, and rank the severity of each error. In coding errors we adopted the categorization scheme used by Foushee, Lauber, Baetge, & Acomb (1986) and later by Wiener, et al. (1991). Crew errors were categorized into three types, I, II, and III:

Type I: *Minor*, with a low probability of serious flight safety consequences. Examples of Type I crew errors include the failure to declare a normal checklist complete, or starting a normal checklist without waiting for the captain's command.

Type II: *Moderately severe*, with a stronger potential for flight safety consequences. Examples of Type II crew errors include declaring an emergency checklist complete without completing the clean-up items, or providing the pilots with approach speeds verbally rather than in writing.

Type III: *Major*, operationally significant crew errors having a direct negative impact on flight safety. Examples of Type III errors include calculating an incorrect pitch attitude (for the Two Engine go-around landing maneuver), or skipping the landing check entirely.

Table 1 below presents the frequencies for each crew error type for both ICIS and CURRENT conditions.

Table 1. Frequency of crew errors by type for each Environment

	Crew Error Type			
ENVIRONMENT	I	II	III	Total
CURRENT	85	80	8	173
ICIS	59	45	1	105
Total	144	125	9	278

Crew error types I, II, and III, and total crew errors, were analyzed using separate t-tests, with ENVIRONMENT (ICIS, CURRENT) as the independent variable.

For Type I and Type III errors, there were no significant difference in mean errors as a function of ENVIRONMENT. For Type II errors, crews using ICIS had significantly fewer errors than did CURRENT crews, $t(1,22)=2.49$; $p<.05$.

When all crew error types were combined, crews using ICIS had significantly fewer total errors than did CURRENT crews, $t(1,22)=2.92$; $p < .01$.

Efficiency

For the purpose of this study, efficiency was considered a multiple-factor construct which included workload, situational awareness, crew resource management (CRM), technical proficiency, compliance with Standard Operating Procedures (SOP), and task execution time. The results for each factor are reported separately.

Workload

Participant response to the NASA Task Load Index (TLX) (Hart & Staveland, 1988) served as the measure of workload.

The ENVIRONMENT x TLX DIMENSION repeated measures Analysis of Variance (ANOVA) revealed both a main effect of ENVIRONMENT, as well as an interaction between ENVIRONMENT and TLX Dimension. For the ENVIRONMENT main effect, the analysis showed that participants who performed the simulator scenario with ICIS had a mean TLX score of 3.17 (possible scale is 1-low to 7-high) which was significantly lower (indicating less workload) than those who performed the scenario without ICIS (score 4.36), $F(1,70)=25.57$; $p<.0001$.

The ENVIRONMENT x TLX DIMENSION interaction was significant, $F(5,350)=6.96$; $p<.00001$. The interaction revealed that not all TLX dimension scores were equally different for CURRENT and ICIS conditions. A post-hoc examination of individual means (Tukey's post-hoc Honest Significant Difference test) showed that scores on four dimensions ("Mental Demand," "Physical Demand," "Temporal Demand," and "Effort") were significantly lower for ICIS condition than for CURRENT; scores on the dimensions "Perceived Performance" and "Frustration" were lower, though not statistically significant.

Situational Awareness

Situational awareness was measured by participant response to the 10 statements of the Situational Awareness Rating Technique (SART) grouped into 3 subscales: Understanding (SART-U), Demand (SART-D), and Supply (SART-S), (Selcon & Taylor, 1989). A composite SART score (SART-Calculated, or SART-C) was then calculated from these 3 subscales:

$$(\text{SART-C})=\text{SART-U} - (\text{SART-D} - \text{SART-S})$$

Each of the SART subscale scores and the SART-Calculated score were analyzed with separate ANOVAs.

The ANOVA for the SART sub-scale Understanding showed no significant differences between CURRENT and ICIS conditions. For the SART Demand sub-scale there was a significant difference, $F(1,70)=20.26$, $p < .0001$, with crews in the CURRENT condition reporting significantly higher Demand than did those in the ICIS condition. There was no significant difference for the SART sub-scale Supply.

The ANOVA performed for SART-Calculated also showed a significant difference for CURRENT and ICIS conditions, $F(1,62)=12.23$; $p<.001$. Participants who used ICIS in the simulator session had a mean SART-C score of 6.4, (higher situational awareness) while those who did not use ICIS in the session had a mean SART-C score of 4.86.

Crew Resource Management (CRM)
There was no significant effect of ICIS on crew resource management.

Crew technical proficiency)
There was no significant effect of ICIS on crew technical proficiency.

Compliance with Standard Operating Procedure
The independent observer rated that ICIS did not interfere with crew compliance with Standard Operating Procedures.

Task Execution Time
During the video analysis, exact times for the start and finish of each procedure were recorded. The analysis of the timing data indicated that there was no significant difference in the total time for the scenario as a function of ICIS or CURRENT condition. Nor was there a significant difference in the execution time for any of the procedures performed during the scenario.

Usability

Questionnaire items (Post-Training and Post-Simulator) were designed to assess participants' opinion on ICIS usability. Item Analysis was used to test the correlation between questionnaire items for each of the two questionnaires; the result of the analysis (Cronbach's alpha > .81 Post-Training) validated the reliability of a composite measure of the Post-Training questionnaire items, but not for the Post-Simulator items. In the Post-Training composite 90% of the participants responded *Strongly Agree* or *Agree* that ICIS is usable. For the Post-Simulator items, 77% of the participants responded *Strongly Agree* or *Agree* that "I can easily find the information I need in ICIS."

Acceptability

Other Post-Training and Post-Simulator questionnaire items assessed participants' opinion on ICIS acceptability. Item Analysis was again used to test the correlation between questionnaire items for each of the two questionnaires; the result of the analysis (Cronbach's alpha > .84 Post-Training, and >.75 Post-Simulator) validated the reliability of composite measures for both questionnaires. Finally, the correlation between the two composite measures was sufficiently high ($r=.48$, $p<.01$), thus indicating that the two measures addressed the same aspects of Acceptability. In the Post-Training composite, 99% of the participants responded *Strongly Agree* or *Agree* that ICIS is acceptable. For the Post-Simulator composite Acceptance measure, 100% of the participants responded *Strongly Agree* or *Agree* that ICIS is acceptable.

DISCUSSION

The ICIS evaluation was presented as an example of a multi-dimensional approach to the assessment of a new cockpit system. Four dimensions provided the framework for the evaluation, enabling a comprehensive examination of the impact of the system on safety, on the efficiency of crew performance, and the degree to which the system was usable and accepted by the crew. Further, the multiple dimensions served to expose any instance in which a positive contribution of the system to one aspect of crew performance was counterbalanced by a negative impact on another. Thus, for example, it was possible to

assess whether reduced errors caused an increase in execution times (a speed-accuracy trade-off), or whether the reduction in workload was accompanied by any decrements in task performance, which might have indicated a 'pilot-out-of-the-loop' effect. Finally, the use of line crews, a full-motion simulator, and a realistic scenario provided assurance that the results would be applicable to a real-life flight situation.

The results of the study provided the designers with a thorough understanding of the impact of the new system on crew performance. Crew errors were significantly reduced, and efficiency, as measured by decreased workload and increased situational awareness, was significantly improved. These findings are perhaps the result of the task-oriented, integrated nature of the information in the ICIS, which enabled immediate access to all the information necessary for optimal performance of each task. This immediacy of information access may have enabled the crew to give their complete attention to the task rather than to searching for and retrieving necessary information. Other efficiency factors such as execution times, crew resource management, technical proficiency, or adherence to standard operating procedure, were not seen to be influenced by the ICIS. This perhaps indicates that the introduction of ICIS did not disrupt the current workflow in the cockpit. Further, the findings showed that the system was usable, in the crews' opinion, and acceptable. The lack of negative findings on any of the dimensions suggests that no trade-off effects were created by the use of ICIS.

CONCLUSIONS

In the introduction to this paper we noted that the need for a formal human factors evaluation methodology is becoming increasingly urgent. Continual improvement in computer technologies (as well as increasingly crowded airways) drives the development of complex, integrated systems with a man-machine interface. Regulatory agencies are required to certify these systems, and to date there is no defined methodology which satisfactorily addresses the human factors aspect.

The evaluation methodology presented here is not suitable for the assessment of every system. As suggested earlier, a number of factors such as system criticality, functionality, and integration with other systems will define the necessary detail of an evaluation. Indeed, an important focus for future work should be the complete definition of the system factors which will determine evaluation needs, and the development of evaluation methodologies for specific system types. However, the multi-dimensional approach, which provides a comprehensive picture of the system within its intended operational environment by using a large number of system-relevant measures, may serve as a useful template for formal methodologies for system development and certification.

ACKNOWLEDGMENTS

The authors would like to thank Ms. Gilly Leshed for her help and support in both phases of the ICIS Evaluation. We would also like to thank the airline, the pilots, and the support personnel who participated in the study.

REFERENCES

Abbott, K., Slotte, S. M., & Stimson, D.K. (1996). *The Interface Between Flightcrews and Modern Flight Deck Systems*, Washington, DC: Federal Aviation Administration.

Avionitek, (1998). *ICIS-Integrated Crew Information System Evaluation Program* (Report AHF-GEN-0001-DO). Eagan, MN. Avionitek, Inc. (available from the author)

Foushee, H.C., Lauber, J.K., Baetge, M.M., & Acomb, D.B. (1986). *Crew factors in flight operations III: The operational significance of exposure to short-haul air transport operations*. Moffett Field, CA: NASA Ames Research Center.

Hart, S.G. & Staveland, L.E. (1988). Development of the NASA-TLX (task load index): Results of empirical and theoretical research. In P.A. Hancock and N. Meshkati (Eds.) *Human mental workload*. Amsterdam, North Holland Press.

RTCA (1999). Final Report of the RTCA Task Force 4: Certification. February 26.

Selcon, S.J. & Taylor, R.M. (1989). Evaluation of the Situational Awareness Rating Technique (SART) as a Tool for Aircrew Systems Design. In *Proceedings of the AGARD AMP Symposium on Situational Awareness in Aerospace Operations*, Copenhagen, DK., 2-6 October, 1989.

Shamo, M.K., Dror, R., & Degani, A. (1998). Evaluation of a new cockpit device: The integrated electronic information system. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, Chicago, Ill.

Small, R.L. & Rouse, W.B. (1994). Certify for Success: A Methodology for Human-Centered Certification of Advanced Aviation Systems. In J. Wise, V.D. Hopkins, and D.J. Garland, (Eds.). *Human Factors Certification of Advanced Aviation Technologies*. Daytona Beach: Embry-Riddle Aeronautical University Press.

Stein, E.S., & Wagner, D. (1994). A Psychologist's View of Validating Aviation Systems. In J. Wise, V.D. Hopkins, and D.J. Garland, (Eds.). *Human Factors Certification of Advanced Aviation Technologies*. Daytona Beach: Embry-Riddle Aeronautical University Press.

Wiener, E.L., Chidester, T.R., Kanki, B.G., Palmer, E.A., Curry, R.E., & Gregorich, S.E. (1991). *The impact of cockpit automation on crew coordination and communication: I. Overview, LOFT evaluations, error severity, and questionnaire data*. Moffett Field, CA: NASA Ames Research Center.